Learning and teaching in repeated games: A machine learning approach to long-term best-response play

D.P. de Farias and N. Megiddo

June 12, 2003

Repeated games and long-term best-response play. We consider the problem of learning during a play of a a repeated game. Existing approaches often focus on algorithms that asymptotically achieve an equilibrium of the one-shot game. We shift away from that paradigm to consider learning strategies that attempt to play long-term best-response to the opponent's observed strategy. This goal may be more appropriate for several reasons:

- **Bounded Rationality.** If the opponent's strategy is not part of an equilibrium, then it is also not necessarily the best for a player to play an equilibrium strategy. Since the game is played repeatedly, there may be an opportunity to *learn* about the opponent's strategy and play an approximate a best-response to it.

- **Multiple Equilibria.** Repeated non-zero-sum games typically possess equilibrium points with payoffs higher than those of the stage game equilibria. This fact suggests that players may be able to *teach* each other to cooperate and trust, in order to converge to a more desirable equilibrium than they could otherwise.

Experts algorithms. For some strategies of the opponent in a repeated game, it is impossible to learn the best-response. In some cases, one may be able to design a relatively small number of strategies, at least one of which can be expected to work well in real-life situations. Therefore, a reasonable question is how to play a repeated game when only a limited number of strategies are available. A related question has been addressed, in a more general setting than repeated games, in the field of machine learning, where so-called "experts algorithms" have been proposed. The goal of these methods is to learn from experience how to combine advice from multiple experts so as to make good sequential decisions on-line. The general idea can be described as follows. A decision maker has to choose repeatedly from a given set of actions. The payoff in each stage is a function of the chosen action and the choices of Nature. A set of strategies $\{1, \ldots, r\}$ is available for the decision maker to choose from. We refer to each such strategy as an "expert," even though some of them might be simple enough to be called novices. Each expert suggests a choice of an action based on the history of the process and the expert's own choice algorithm. After each stage, the decision maker observes his own payoff. An experts algorithm directs the decision maker with regard to which expert to follow in the next stage, based on the past history of actions and payoffs.

Minimum Regret. Minimum Regret (MR) is a popular criterion in decision making processes. Regret is defined as the difference between the payoff that could have been achieved, given the choices of Nature, and what was actually achieved. An expert selection rule is said to minimize regret if it yields an average payoff as large as that of any single expert, against any fixed sequence of actions chosen by the opponent. Indeed, certain experts algorithms have been shown to minimize regret. Such algorithms choose at each stage an expert from a probability distribution that is related to the payoff accumulated by the expert prior to that stage. It is crucial to note though that, since the experts are compared on a sequence-by-sequence basis, the MR criterion ignores the possibility that different experts may induce different sequences of choices by the opponent. Thus, MR makes sense only under the assumption that Nature's choices are independent of the decision maker's choices.

In repeated games, the assumption that the opponent's choices are independent of the player's choices is not justified, because the opponent is likely to base his choices of actions on the past history of the game. This is evident in nonzero-sum games, where players are faced with issues such as how to coordinate actions, establish trust or induce cooperation. These goals require that they take each other's past actions into account when making decisions. But even in the case of zero-sum games, the possibility that an opponent has bounded rationality may lead a player to look for patterns to be exploited in the opponent's past actions.

We illustrate some of aforementioned issues with an example involving the Prisoner's Dilemma game.

The Prisoner's Dilemma. Suppose in the repeated Prisoner's Dilemma game row player consults with a set of experts, including the "defecting expert," who recommends defection all the time. Let the strategy of the column player in the repeated game be fixed. In particular, the column player may be very patient and cooperative, willing to wait for the row player to become cooperative, but eventually becoming non-cooperative if the row player does not seem to cooperate. Since defection is a dominant strategy in the stage game, the defecting expert achieves in each step a payoff as high as any other expert against any sequence of choices of the column player, so the row player learns with the experts algorithm to defect all the time. Obviously, in retrospect, this seems to minimize regret, since for any fixed sequence of actions by the column player, constant defection is the best response. Obviously, constant defection is not the best response in the repeated game against many possible strategies of the column player. For instance, the row player would regret very much using the experts algorithm if he were told later that the column player had been playing a strategy such as Tit-for-Tat.

Talk outline. We frame the problem of learning in games from a machine learning perspective, and offer an introduction to fundamental ideas and results in the field that may be relevant to this problem. We then describe our preliminary results involving a new experts algorithm, which follows experts judiciously, attempting to maximize the long-term average payoff. Our algorithm differs from previous approaches in at least two ways. First, each time an expert is selected, it is followed for multiple stages of the game rather than a single one. Second, our algorithm takes into account only the payoffs that were actually achieved by an expert in the stages it was followed, rather than the payoff that could have been obtained in any stage. Our algorithm enjoys the appealing simplicity of the previous algorithms, yet it leads to a qualitatively different behavior and improved average payoff. We present two results:

1. A "worst-case" guarantee that, in any play of the game, our algorithm achieves an average payoff that is asymptotically as large as that of the expert that did best in the rounds of the game when it was played. The worst-case guarantee holds without any assumptions on the opponent's or experts' strategies.

2. Under certain conditions, our algorithm achieves an average payoff that is asymptotically as large as the average payoff that could have

been achieved by the best expert, had it been followed exclusively. The conditions are required in order to facilitate learning and for the notion of a “best expert” to be well-defined.

The effectiveness of the algorithm is demonstrated by its performance in the repeated PD game, namely, it is capable of identifying the opponent’s willingness to cooperate and it induces cooperative behavior.